Do linguists underestimate the extent of gradience in syntactic data?

Introduction Many syntactic theories assume that grammaticality is a binary distinction, though there are exceptions, e.g. Linear Optimality Theory (Keller 2000). Gradience only plays a minor role in syntax: While a vast majority of papers in *Linguistic Inquiry* (2001-2010) apply a gradient scale with more than two levels, only few items were judged as “in-between” (for details see below). (On the other hand, it has long been noted that perceived acceptability is a gradient phenomenon, but this is typically attributed to extra-grammatical factors.)

Is this prevalence of unmarked and *-marked items evidence for a sharp division between a class of grammatical sentences and a class of ungrammatical sentences as argued by Chomsky (1955/75) and many others? Or do linguists consistently underestimate the extent of gradience? To address this question, we randomly sampled 200 sentences from LI articles and had non-linguists rate them.

Corpus Similar to Spouse et al. 2013, we created a corpus of acceptability judgments given by linguists in articles published in LI 2001-2010. Notably, most papers distinguish more than two levels of acceptability: 2133 out of 2619 items (81%) involving a standard acceptability judgment come from papers which use “?”, “?*”, etc. to indicate intermediate levels of acceptability. At the same time, only 157 items (6%) are marked by some diacritic other than “*”, whereas 969 items are marked by an asterisk (37%) and 1493 items are unmarked (57%).

From the corpus, we randomly selected 200 items in four categories: 100 items from papers which include intermediate levels of acceptability ("gradient papers") and 100 items from papers with only two levels of acceptability ("binary papers"). Within each set, we extracted 50 items which are unmarked ("OK-items") and 50 items which are marked by an asterisk ("*-items").

Experiment We ran two experiments. In Experiment 1, we obtained acceptability ratings from 80 participants using a 7-point scale, in Experiment 2, we obtained binary acceptability rating for the same sentences from another 80 participants. We used Amazon Mechanical Turk to recruit subjects and a separate website to actually run the experiments. 15 participants in Experiment 1 and 12 participants in Experiment 2 were excluded, based on the following criteria: not being a native speaker of American English (determined post-experimentally), returning incomplete results, having extreme reaction times, or failing on “booby trap” items.

Overall, experimental results and LI ratings correlate to a reasonable extent (correlation coefficients range from .65 to .77). Experimental ratings only rarely contradict author ratings flat out. Yet, we find important discrepancies: As illustrated in Figure 1, mean ratings in Experiment 1 cover the whole space on the experimental 7-point scale rather than clustering at the two extremes (as one would expect given that we extracted OK-items and *-items). Likewise, acceptance rates in Experiment 2 range from 0 to 1 with no gap in-between (see Figure 2).

Instead of the S-curve that one would expect as a noisy approximation of the step-function corresponding to a sharp binary distinction, the graphs show a steady increase from 1 to 7 (for Experiment 1) and from 0 to 1 (for Experiment 2). This finding is particularly striking for items from gradient papers: One would expect not a single item in the intermediate range, because we restricted our item selection to items that the LI authors either marked as ungrammatical (*-items) or left them unmarked (OK-items) (despite their use of intermediate levels in their papers). Fisher's Exact Tests indicate that the observed number of in-between items (43 in Experiment 1 and 25 in Experiment 2) is significantly higher than the expected number of in-between items (0) in both experiments (p <.001; we did not test binary papers, as we would have to speculate about the expected number of in-between items).

In both experiments and both for items from gradient and binary papers, the intermediate range is densely populated. Notably the number of in-between items is higher in Experiment 1 than in Experiment 2 and higher for items from gradient papers compared to items from binary papers. (NB: Interestingly, most in-between items from gradient papers are *-marked in the corresponding paper.)

Discussion One might object that the observed gradience is an experimental artefact due to aggregation. However, ratings in Experiment 1 argue against this as they cover the entire scale; and they cluster in the intermediate range both at the level of items means and at the level of individual ratings. Another objection could be that gradience in acceptability is all due to performance factors while the underlying grammaticality is binary, as argued inter alia by Newmeyer (2003). We agree that performance factors affect acceptability ratings. The current data set, however, does not involve any garden-path sentences, multiple center embedding. We therefore believe that known performance factors cannot account for the prevalence of intermediate values nor for the shape of the curve. Further, the fact that most items with intermediate ratings were judged ungrammatical in LI also argues against attributing all gradience to performance. (NB: Though performance factors can also increase acceptability ratings and even create illusory grammaticality, cf. Phillips et al. 2011, it is unlikely that they are the driving force in the current data set which does not exhibit any resemblance to know grammaticality illusions. Further, the observed increase cannot be due to repair readings, either, as we excluded sentences with possible repair readings.)

We conclude that the observed gradience is not an epiphenomenon or artefact. Moreover, our data show that gradience is far more prevalent than previously thought. This should be reflected in theory building.

References

Chomsky, N. (1955/1975). *The logical structure of linguistic theory*. Chicago: University of Chicago Press.

Keller, F. (2000). Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality. Ph.D. thesis, University of Edinburgh.

Newmeyer, F. J. (2003). Grammar is grammar and usage is usage. *Language 79*, 682-707.

Phillips, C., Wagers, M.W., and Lau, E.W. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. In J.T. Runner (ed.), *Experiments at the Interfaces*, 153-218. Bingley: Emerald.

Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. *Lingua 134*, 219-248.

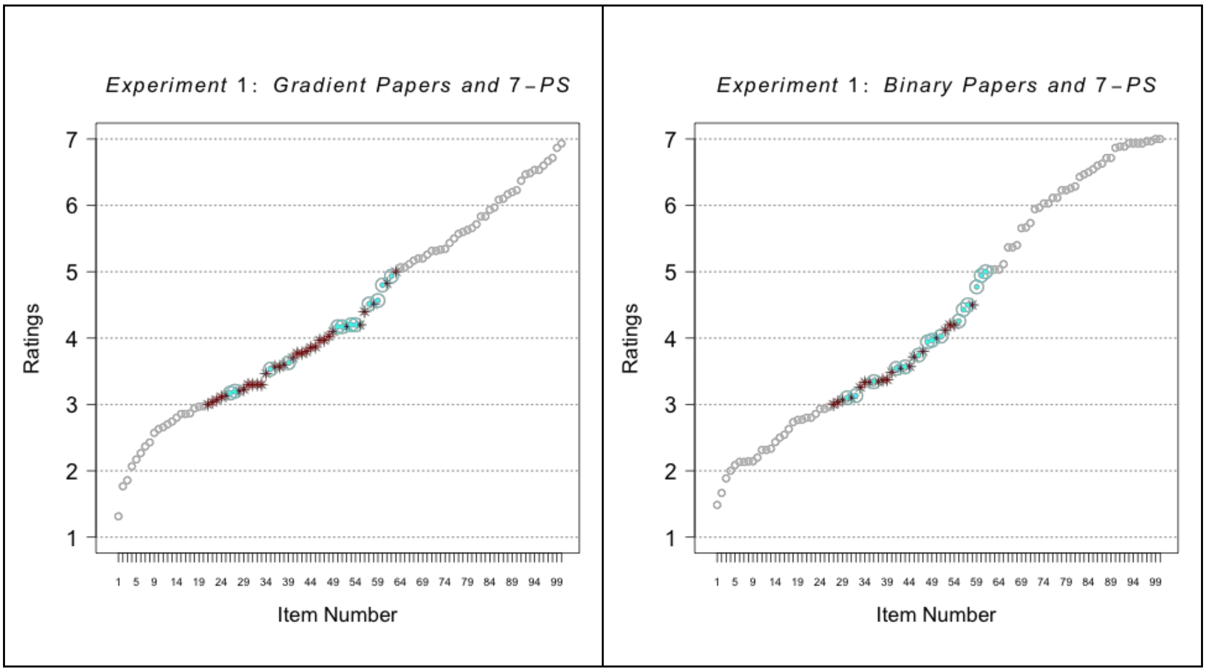

Figure 1. Item means in Experiment 1 for the set of items from gradient papers (left) and for the set of items from binary papers (right). Items with a mean rating >3 and <5 (gradient papers: 43 items, binary papers: 35 items) are color-marked (blue: OK-items, red: *-item).

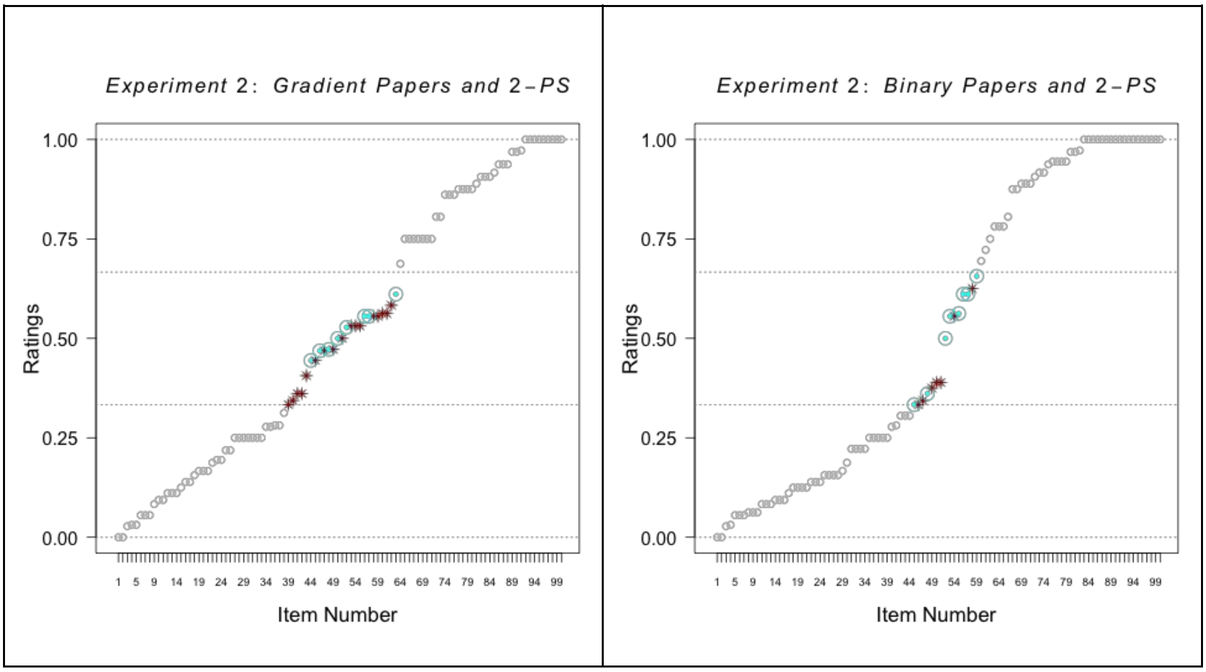

Figure 2. Mean acceptance rates in Experiment 2 for items from gradient papers (left) and for items from binary papers (right). Items with acceptance rates >.33 and <.66 (gradient papers: N=25, binary papers: N=15) are color-marked (blue: OK-items, red: *-items).